URL Frontier

GOALS

Large scale crawl using StormCrawler and URLFrontier

Check that it could cope with large volume of data

In a timely fashion

Fix bugs and improve performance

Generate web archives to be published by CommonCrawl

CHALLENGES

Totally new piece of software

Running on generic hardware

Sending large amount of data to AWS S3

DEMO SETUP

Virtual Wall (IMEC)

6 nodes in total pcgen04

5 worker nodes (doing the crawling)

1 master node running URL Frontier

RESULTS

fetched 354M URLs

1.2B URLs discovered but not yet fetched

Total size of the WARC files on AWS S3 was 36.8 TB

MORE RESULTS

Pushed single URLFrontier instance to its limits

Great success nonetheless!

Fixed quite a few bugs

Added substantial performance improvements

CONCLUSIONS

Very successful experiment

Public dataset not yet available

Validated our technical choices

POST MORTEM

Great performance on a single Frontier instance

Helped think about improvements

Work as a cluster

New NLNet project URLFrontier2

Under way

Already have an organisation in stealth mode evaluating it !